AFRL-HE-WP-TP-2006-0040

AIR FORCE RESEARCH LABORATORY

Psychometric Correlates of the Effects of Image-Enhancing Algorithms on Visual Performance

George Reis
Alan Pinkus
Kelly Neriani

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

February 2006

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2006	Technical Paper	

4. TITLE AND SUBTITLE
Psychometric Correlates of the Effects of Image-Enhancing Algorithms On Visual Performance

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
George Reis, Alan Pinkus, Kelly Neriani

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
11

5f. WORK UNIT NUMBER
27

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECV

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-WP-TP-2006-0040

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This will be published in the Proceedings of the SPIE Defense & Security Symposium Conference. The clearance number is AFRL/WS-06-0613, cleared 6 March 2006.

14. ABSTRACT
Future military image acquiring devices will have computational capabilities that will allow agile, realtime image enhancement. In preparing for such devices, numerous image enhancement algorithms should be studied; however, these algorithms need evaluating in terms of human visual performance using military-relevant imagery. Evaluating these algorithms through objective performance measures requires extensive time and resources. We investigated subjective methods for down-selecting algorithms to be studied in future research, and thus, provide a methodology for down-selection, Imagery was processed using six algorithms and then ranked along with two baselines through the method of paired comparisons and the method of magnitude estimation, in terms of subjective attitude. These rankings were then compared to objective performance measures: reaction times and errors in finding targets in the processed imagery. In general, we found associations between subjective and objective measures. This leads us to believe that subjective assessment may provide an easy and fast way for down-selecting algorithms but at the same time should not be used in place of objective performance-based measures.

15. SUBJECT TERMS
Image enhancement, Method of paired comparisons, Magnitude estimation, Psychometrics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			George Reis
UNC	UNC	UNC	SAR	10	19b. TELEPHONE NUMBER *(include area code)* (937) 255-8863

Psychometric correlates of the effects of image-enhancing algorithms on visual performance

George A. Reis[a], Kelly E. Neriani[a,b], Alan R. Pinkus[a], Eric L. Heft[a]

[a]Air Force Research Laboratory AFRL/HECV, 2255 H St., Wright-Patterson AFB OH, USA 45433
[b]Consortium Research Fellows Program, 2511 Jefferson Davis Highway, Arlington, VA, USA 22202-3926

ABSTRACT

Future military image acquiring devices will have computational capabilities that will allow agile, realtime image enhancement. In preparing for such devices, numerous image enhancement algorithms should be studied; however, these algorithms need evaluating in terms of human visual performance using military-relevant imagery. Evaluating these algorithms through objective performance measures requires extensive time and resources. We investigated subjective methods for down-selecting algorithms to be studied in future research, and thus, provide a methodology for down-selection. Imagery was processed using six algorithms and then ranked along with two baselines through the method of paired comparisons and the method of magnitude estimation, in terms of subjective attitude. These rankings were then compared to objective performance measures: reaction times and errors in finding targets in the processed imagery. In general, we found associations between subjective and objective measures. This leads us to believe that subjective assessment may provide an easy and fast way for down-selecting algorithms but at the same time should not be used in place of objective performance-based measures.

Keywords: image enhancement, method of paired comparisons, magnitude estimation, psychometrics

1. INTRODUCTION

1.1 Image Enhancement

As night vision devices continue to provide us with advantages in military night operations, there are new related capabilities that are emerging and being explored. These new capabilities involve the realtime enhancement of as-acquired imagery from night vision and camera devices in compact computing hardware.

Image enhancement allows the increase in saliency of a certain aspect or component (e.g., contrast) of the image makeup and thus increases the saliency of an element or object (e.g., tank) in the image. In military-relevant imagery, such an allowance may assist operators in finding objects faster and in identification of objects. As the benefits of multi-sensor fusion emerge, image enhancement research in this field is also making headway. For example, Rahman, Jobson, Woodell, & Hines[1] presented an approach to sensor fusion and enhancement using the Multiscale Retinex image enhancement algorithm. The results of such studies may help pilots land their aircraft in poor visibility conditions. In addition to military contexts, image enhancement may increase performance in other fields. The range includes law enforcement and homeland security (e.g., surveillance), forensics (e.g., fingerprint enhancement), and medical imaging (e.g., chromosome karyotying, lesion detection). Although image enhancement research is conducted in the areas mentioned above, there does not appear to be a vast amount of research in objective methods for assessing visual performance-enhancing effects.

1.2 Assessment

In image enhancement research, numerous studies describe quality metrics. These measures are derived mathematically and perceptually. What is sparse in the literature is research that describes performance related effects of enhanced imagery. For example, in military-relevant imagery, operators may try to find objects (e.g., vehicles, combatants, instillations), and if this imagery is degraded (e.g., sensor failure, weather interference, fidelity limitations),

image enhancing algorithms may be applied. It would be desirable to know which algorithm allows the operator to find the target object quickest and with the best accuracy. One such study performed by Neriani, Herbranson, Pinkus, Task & Task[2] measured response time and percent correct in enhanced images versus baseline, non-enhanced images. Their study discusses some of the intricate issues that researchers should consider in the context of enhanced imagery of vegetated terrain.

Many variables can add intricacy to image enhancing research. To add, performance measures require many resources such as time, experimental participants, computing power, and mass amounts of imagery across different sensors, weather conditions, terrains, and targets. Moreover, thousands of algorithms and algorithm parameters can be chosen for study. How does one go about down-selecting algorithms? This is the crux of the study presented here. We propose a method for down-selecting algorithms by subjective assessment of the imagery that would be used in studies that utilize performance-based measures. We hypothesize that associations exist between subjective assessment, as measured through psychometric methods, and performance-based assessment.

Our hypothesis is based on studies of image quality and intuitiveness. As for intuitiveness, we ask ourselves, "If I think this image will allow me to find a target faster, then should I have a shorter response time in actually finding the target?" Our intuition tells us *yes*. As for the other basis in our hypothesis, we look at image quality research. Zhou & Bovik[3] observed consistent correlation between subjective measures and their proposed universal image quality index. Leachtenaurer[4] also observed correlations between objective quality measures and analyst ratings. Image quality may be related to the intricacies of our exploration of associations between subjective and objective performance measures, but it should be clear that there is a difference between what *looks good* and what *facilitates good performance*. To help us determine if certain algorithm-enhanced imagery facilitates good performance, in terms of subjective measures, we have employed two psychometric methods: the method of paired comparisons and magnitude estimation.

1.3 Psychometric Measures

In trying to discover associations between subjective measures and objective performance-based measures, we used the methods of paired comparisons and magnitude estimation independently of each other. There are other psychometric methods that could have been used, but these two methods provide an easy and fast method for data collection, which, in essence, is the core issue of this study—determining a fast, easy, low-resource method for down-selecting algorithms to test. Other psychometric methods should be tested as well for they may provide additional understanding in the subject matter.

In the method of paired comparisons, participants are shown pairs of images and are asked to choose which of the two is preferred based on some characteristic. This method uses the number of times each image is preferred over another across subjects and trials as the scale value. This method was used in this study due to its simplicity and ability to compare the preference patterns of individual participants.

In magnitude estimation, participants are asked to rate test items against a reference item that is given some numerical base value. If participants wish to rate the test item twice as high as the reference item on some particular characteristic then they would rate the test item twice as much as the base value. If the test item is perceived as being half as much as the reference item, then participants would rate the test item half the base value.

1.4 Objective Performance-based Measures

To complete this study, we needed subjective and performance-based measures to explore possible associations. The subjective measures collected in this study corresponded to the performance-based measures obtained by Neriani, Herbranson, Reis, Pinkus, and Goodyear[5]. This study utilized participants and imagery from the Neriani et al.[5] study for comparability. Neriani et al.[5] assessed six algorithms that are specifically designed to enhance the contrast of digital images. The image enhancing algorithms used in this study included the Multiscale Retinex (MSR) algorithm, Block-based Binomial Filtering Histogram Equalization (BBFHE), Global Histogram Equalization (HE), Partially Overlapped Sub-block Histogram Equalization (POSHE), the Autolevels function, and the Recursive Rational Filter (RRF) technique. For a summary explanation of these algorithms, see Neriani et al.[5] Neriani et al.'s[5] method describes the acquiring of objective human visual performance data as a means of evaluating contrast enhancement algorithms. Their approach uses standard objective performance metrics, such as response time and error rate, to compare algorithm-enhanced images versus two baseline conditions, original non-enhanced images and contrast-degraded images. Observers completed a visual search task using a spatial-forced-choice paradigm. Observers had to search images for a

target (a military vehicle) hidden among foliage and then indicate in which quadrant of the screen the target was located. Response time and percent error were measured for each observer.

2. METHODS

2.1 Participants

Fourteen observers, ten males and two females, participated in this experiment. The observers ranged in age from 21 to 50 years. All had normal color vision and normal or corrected-to-normal visual acuity. Seven of the 14 participants tested in both the Neriani et al.[5] study and this study. These participants will be referred to as belonging to the experienced group. The other seven (not tested in Neriani et al.[5]) will be referred to as belonging to the inexperienced group.

2.2 Stimuli

The imagery used in this study was obtained from that used in Neriani et al.[5] Although a very small subset of imagery from their study is utilized in this study, the results in this study are compared to all the images used in their study.

In Neriani et al.[5], each observer viewed a total of 1696 grayscale images. Of these, 1536 images consisted of a target located in a scene of trees and grass. The target was a model tank, a Renault R39 reconnaissance tank (see Figure 1), placed on an artificial terrain board. For the rest of the 1696 test images, the observers viewed 160 images, as catch trials, of an artificial terrain board scene of trees and grass with no target.

All 1696 images were presented in one of eight different algorithm conditions. These conditions were a non-degraded condition with no algorithm, a contrast-degraded condition with no algorithm, and six conditions corresponding to contrast-degraded images processed by each of the six contrast-enhancing algorithms described above. The images for the non-degraded condition with no algorithm were taken with the digital camera set to have an exposure of 1 second, with the camera one meter away from the terrain board, illuminated with floodlights (see Figure 2). The images for the contrast-degraded condition with no algorithm were the same as non-degraded condition with no algorithm images except captured with an exposure of 125 milliseconds. Figure 3 shows the eight different algorithm conditions applied to one scene (be aware of differences between images displayed on the CRT and images displayed on paper). The arrow in the top-left image highlights the target, which is the placed in the same location for each image in Figure 3.

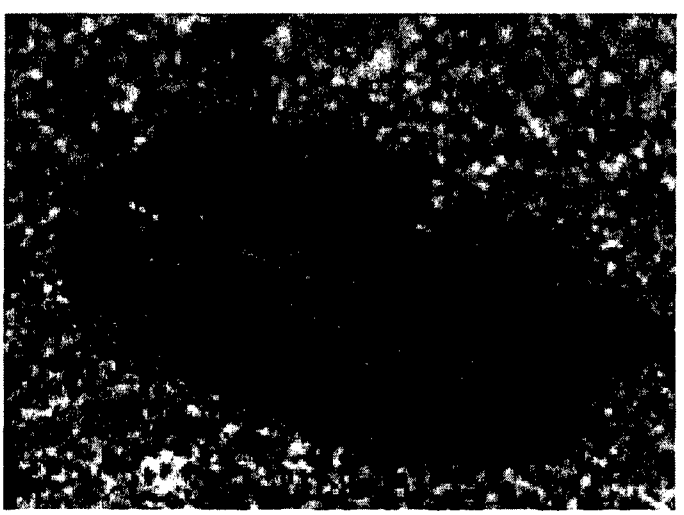

Figure 1. The target used in the images.

Figure 2. The terrain board.

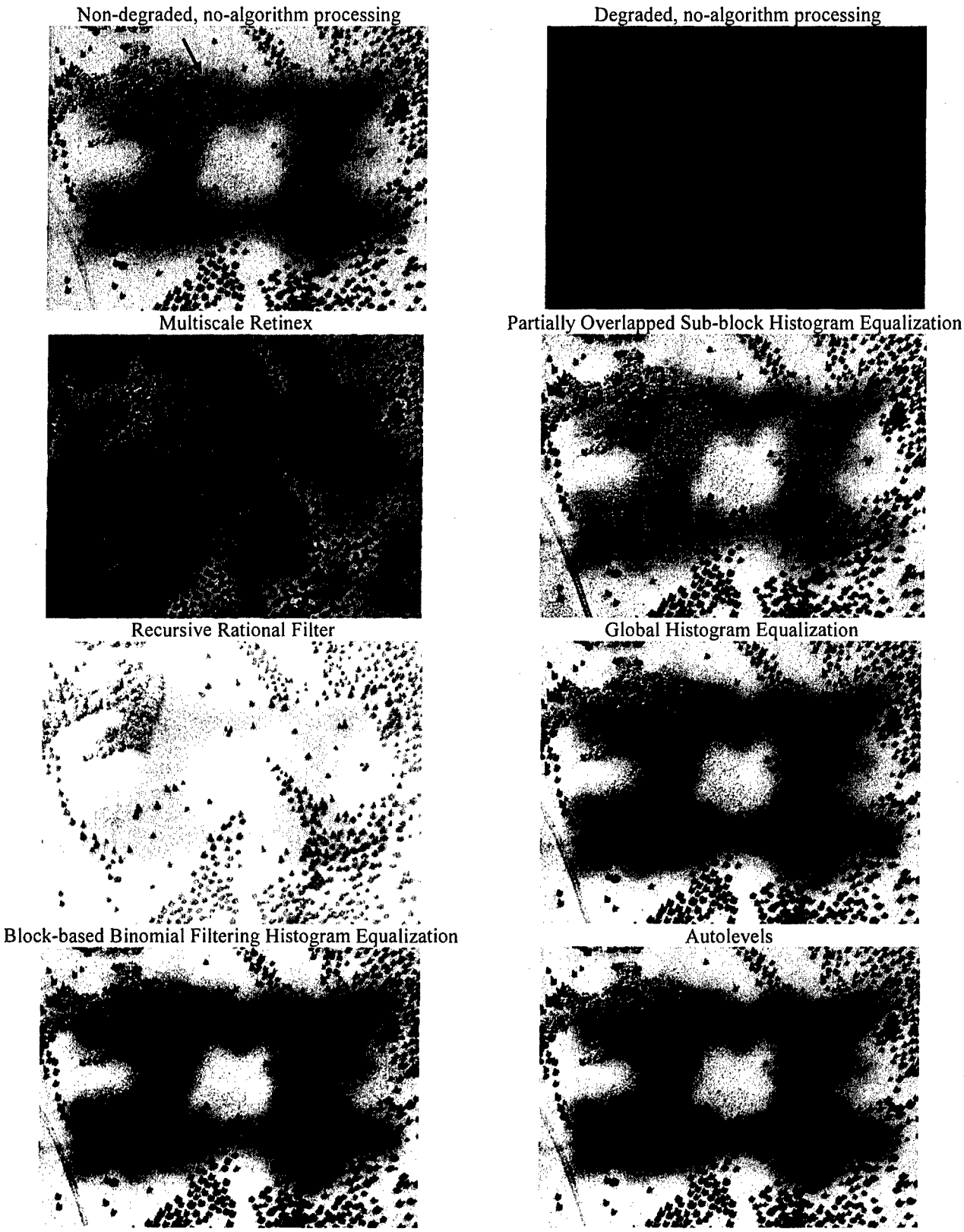

Figure 3. The eight different algorithm conditions. The arrow in top-left image highlights the target.

The images were taken using a Nikon COOLPIX E8800 8.0 megapixel digital camera with a resolution of 1280 x 960 pixels. Before the images were used in the experiment, both the enhanced and non-enhanced images were resized to a resolution of 1155 x 866 pixels. This was necessary due to the constraints of the program running the experiment. For more information about the images, see Neriani et al.[5]

Of the 1536 images (images with targets), three were chosen to represent Neriani et al.[5]'s set in this study. The three images chosen were an image with one of the fastest response times, an image with one of the slowest response times, and one image that ranked medium in response time. The fast response scene will be referred to as the "easy" scene and the slow response time scene will be referred to as the "difficult" scene (see Figure 4). The medium response time image was used for training and the other two were used for analysis. Although the scenes were not intentionally created such that one scene would be more difficult than an other, it can be seen in the images as to why it may have been easier to find the target in one scene over the other. The heading position of the target relative to the scene and possibly the location placement on the terrain board may have contributed to this difference. This difference does not take away from Neriani et al.[5] considering that each scene was tested at each algorithm condition.

2.3 Procedure

Both groups of participants were tested on the method of paired comparisons and magnitude estimation. Participants were tested in paired comparisons first and then in magnitude estimation.

In the paired comparisons test, there was a total of 28 different pairs that were tested (each of the eight algorithms conditions paired against each other). Participants were tested in three blocks of 28 trials each. The first block served as training and familiarization with the program. The participants subsequently were tested in two more blocks. One block utilized the fast response image. The other utilized the slow response image. In each trial, participants were presented with one of the algorithm conditions and then toggled between another algorithm condition until they chose one based on the following instruction: "Pick the image that you think would allow you to find the target faster if you did not know where it was located." The participants were told where the target was in the scene and were told to look at both the target and the entire scene to make their judgment.

After the participants completed the paired comparisons, they were tested in three blocks of magnitude estimation trials. Each block contained eight trials. Each of the eight algorithm conditions was compared to a reference image, the non-degraded no-algorithm condition. The first block served as training and familiarization with the program. The participants subsequently were tested in two more blocks. One block utilized the fast response image. The other utilized the slow response image. In each trial, participants were presented with the non-degraded no-algorithm condition first. The computer display presented a "100" in the top-left corner of the scene. The participants toggled between this image and another image (another algorithm condition), including that of itself. In this second image, an empty field resided where the "100" was in the reference image. Participants were told the following on each magnitude estimation trial: "For the second scene, assign a numerical value that represents how fast you would be able to locate the target in the scene with reference to the first scene having a value of 100." The participants were told where the target was in the scene and were told to look at both the target and the entire scene to make their judgment.

The participants could toggle back and forth as much as they liked in any trial of either psychometric method. The order of easy and difficult response time blocks was counterbalanced across participants. The order of which algorithm condition appeared first in the paired comparisons was randomized across 28 trials for each combination of block and participant. The order in which algorithm condition was presented in the magnitude estimation was randomized across eight trials for each combination of block and participant.

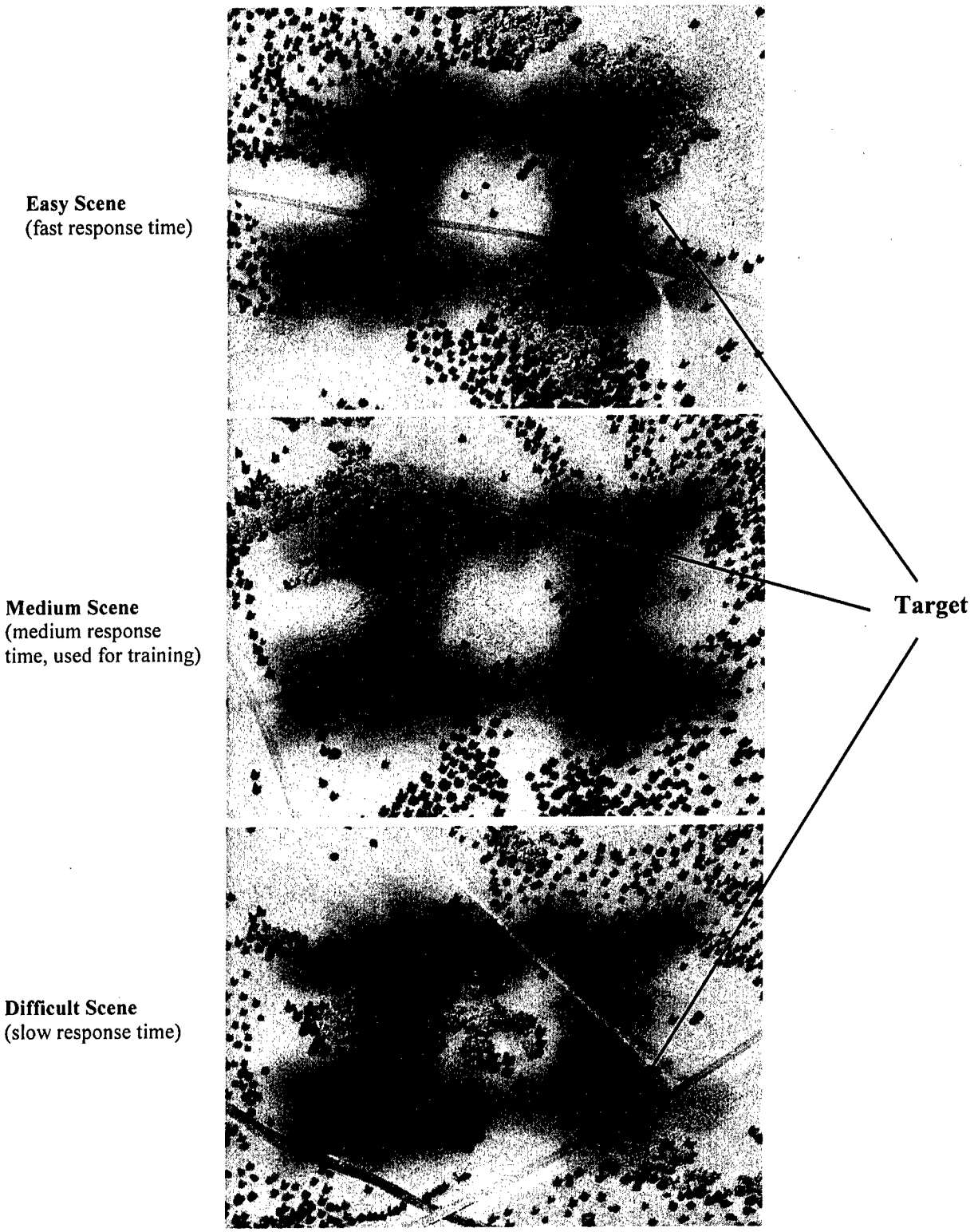

Figure 4. The three scenes used in this study were qualified as easy, medium, and difficult. The easy scene had the fastest response time in trying to find the target. The difficult scene had slowest response time in trying to find the target. The medium scene had a response time that ranked in the middle of response times in trying to find the target. The medium scene was used for training.

3. RESULTS

Our objective was to obtain subjective measures and explore any correspondences with performance-based measures. In the paired comparisons two forced-choice method, we collected such subjective data, but before we proceeded in comparing to the performance-based measures, we analyzed for participant agreement and consistency in responses. This was accomplished by measuring Kendall and Babington Smith's[6] *coefficient of concordance* and *coefficient of consistence*. Through the paired comparisons testing we can obtain an ordering or rank of the algorithm conditions. By performing these analyses, we may gain insight as to whether participants can actually obtain this ordering (i.e., if participants are able to perform the task of discrimination or if participants are possibly not motivated).

The coefficient of concordance is a measure of association that evaluates of the degree of agreement between m sets of ranks for n participants/objects. An ordering or rank of the algorithm conditions is created after all pairs are tested. In analysis, the null hypothesis is H_o: the correlations between the m sets of ranks equals 0. The range of possible values the coefficient of concordance spans is $0 \leq \tilde{W} \leq +1$. When there is complete agreement among all m participants, the value of $\tilde{W}$ is 1. When there is no agreement among the m participants, $\tilde{W}$ is 0.

The coefficient of consistence is used for determining object scalability and individual judge consistency when using complete paired comparison data. A circular triad is formed whenever an inconsistency in pair wise choices occurs. For example, if a participant is presented all pairs of three objects, A, B, and C, and is asked to judge which in the pair is preferred, then a preference pattern of the following type may result:

$$A > B,\ B > C,\ C > A \qquad \text{where} > \text{represents "is preferred over."}$$

This pattern of inconsistency is called a circular triad. Here, the null hypothesis is H_o: preferences among the images are random. If the images were very similar (in the characteristic tested), then one would expect few if any subjects showing a significant test result. In general, the coefficient of consistence is:

$$\zeta = \frac{(\text{number of possible circular triads}) - (\text{number of circular triads})}{(\text{number of possible circular triads})} \qquad \text{where } \zeta \text{ ranges from 0 to 1.}$$

Table 1 shows the computed coefficients of concordance and the corresponding significance levels for each of the scenes (Easy, Difficult) by each group (Experienced, Inexperienced). Table 2 shows the computed coefficients of consistence and the corresponding significance levels for each participant in each combination of scene and group. From Table 1, we observe that there was participant agreement in each of the scene by group combinations. From Table 2, we observe that each of the participants successfully ordered the algorithm conditions. Although some triads existed across the participants, there were not enough to determine that choices were made at random. The participants successfully ordered the algorithm conditions. Since we determined that our paired comparison data was sound, we continued to explore the data for possible associations between subjective measures and the objective performance-based measures.

Table 1. Coefficient of concordance for scene x group.

Scene / Group	$\tilde{W}$	p
Easy / Experienced	0.5596	= 0.0003
Easy / Inexperienced	0.6351	= 0.0001
Difficult / Experienced	0.7314	< 0.0001
Difficult / Inexperienced	0.6935	< 0.0001

Table 2. Coefficient of consistence for scene x group.

Scene / Group = Easy / Experienced		
Participant	ζ	p
1	0.9500	0.0004
2	0.9000	0.0007
3	1.0000	0.0002
4	1.0000	0.0002
5	0.8500	0.0071
6	1.0000	0.0002
7	0.9000	0.0007

Scene / Group = Difficult / Experienced		
Participant	ζ	p
1	1.0000	0.0002
2	1.0000	0.0002
3	1.0000	0.0002
4	1.0000	0.0002
5	1.0000	0.0002
6	0.9000	0.0007
7	1.0000	0.0002

Scene / Group = Easy / Inexperienced		
Participant	ζ	p
1	0.8500	0.0071
2	0.9500	0.0004
3	0.9500	0.0004
4	0.9000	0.0007
5	1.0000	0.0002
6	1.0000	0.0002
7	0.9500	0.0004

Scene / Group = Difficult / Inexperienced		
Participant	ζ	p
1	0.9500	0.0004
2	1.0000	0.0002
3	1.0000	0.0002
4	0.9500	0.0004
5	1.0000	0.0002
6	0.9500	0.0004
7	1.0000	0.0002

By computing Pearson correlation coefficients, we assessed the relationship between the preference scores of the paired comparisons and the magnitude estimation values, separately, with the performance-based measures, response time and percent error, separately. Figure 5 shows the scatter plots of the data obtained from the paired comparisons plotted against response time and percent error. Along the x-axis is the average response time and percent error for all the images used in the objective performance-based test. The paired comparison data is shown as the average number of total *votes* across participant, where in each participant, the total vote is the sum of occurrences where one algorithm condition was chosen over another. Figure 5 also shows scatter plots of the average magnitude estimation values across participants against response time and percent error. An analysis of variance concluded that the groups were not statistically significantly different from one another; thus, the values in the subjective measures are averages across the experienced and inexperienced groups. In each plot, letters that represent the algorithm conditions are superimposed on the graph point they represent and the regression lines for each data set are drawn. The letters and the algorithm conditions they represent are: O = non-degraded, C = RRF, R = MSR, A = AutoLevels, B = BBFE, H = Histogram Equalization, D = degraded, P = POSHE.

In Figure 5, the r^2 value and the p-value for the correlation coefficient for each group x scene condition are shown in the upper right-hand corner of each plot. For the difficult scene, there were significant correlations, $p \leq 0.0109$. As response time and percent error increase, we observe higher magnitude estimation values and higher number of votes in paired comparisons. Although the trend was the same for the easy scene, there were no significant correlations, $p > 0.1105$.

3. RESULTS

Our objective was to obtain subjective measures and explore any correspondences with performance-based measures. In the paired comparisons two forced-choice method, we collected such subjective data, but before we proceeded in comparing to the performance-based measures, we analyzed for participant agreement and consistency in responses. This was accomplished by measuring Kendall and Babington Smith's[6] *coefficient of concordance* and *coefficient of consistence*. Through the paired comparisons testing we can obtain an ordering or rank of the algorithm conditions. By performing these analyses, we may gain insight as to whether participants can actually obtain this ordering (i.e., if participants are able to perform the task of discrimination or if participants are possibly not motivated).

The coefficient of concordance is a measure of association that evaluates of the degree of agreement between m sets of ranks for n participants/objects. An ordering or rank of the algorithm conditions is created after all pairs are tested. In analysis, the null hypothesis is H_o: the correlations between the m sets of ranks equals 0. The range of possible values the coefficient of concordance spans is $0 \leq \tilde{W} \leq +1$. When there is complete agreement among all m participants, the value of $\tilde{W}$ is 1. When there is no agreement among the m participants, $\tilde{W}$ is 0.

The coefficient of consistence is used for determining object scalability and individual judge consistency when using complete paired comparison data. A circular triad is formed whenever an inconsistency in pair wise choices occurs. For example, if a participant is presented all pairs of three objects, A, B, and C, and is asked to judge which in the pair is preferred, then a preference pattern of the following type may result:

$$A > B, B > C, C > A \qquad \text{where} > \text{represents "is preferred over."}$$

This pattern of inconsistency is called a circular triad. Here, the null hypothesis is H_o: preferences among the images are random. If the images were very similar (in the characteristic tested), then one would expect few if any subjects showing a significant test result. In general, the coefficient of consistence is:

$$\zeta = \frac{\text{(number of possible circular triads)} - \text{(number of circular triads)}}{\text{(number of possible circular triads)}} \qquad \text{where } \zeta \text{ ranges from 0 to 1.}$$

Table 1 shows the computed coefficients of concordance and the corresponding significance levels for each of the scenes (Easy, Difficult) by each group (Experienced, Inexperienced). Table 2 shows the computed coefficients of consistence and the corresponding significance levels for each participant in each combination of scene and group. From Table 1, we observe that there was participant agreement in each of the scene by group combinations. From Table 2, we observe that each of the participants successfully ordered the algorithm conditions. Although some triads existed across the participants, there were not enough to determine that choices were made at random. The participants successfully ordered the algorithm conditions. Since we determined that our paired comparison data was sound, we continued to explore the data for possible associations between subjective measures and the objective performance-based measures.

Table 1. Coefficient of concordance for scene x group.

Scene / Group	$\tilde{W}$	p
Easy / Experienced	0.5596	= 0.0003
Easy / Inexperienced	0.6351	= 0.0001
Difficult / Experienced	0.7314	< 0.0001
Difficult / Inexperienced	0.6935	< 0.0001

Table 2. Coefficient of consistence for scene x group.

Scene / Group = Easy / Experienced			Scene / Group = Difficult / Experienced		
Participant	ζ	p	**Participant**	ζ	p
1	0.9500	0.0004	1	1.0000	0.0002
2	0.9000	0.0007	2	1.0000	0.0002
3	1.0000	0.0002	3	1.0000	0.0002
4	1.0000	0.0002	4	1.0000	0.0002
5	0.8500	0.0071	5	1.0000	0.0002
6	1.0000	0.0002	6	0.9000	0.0007
7	0.9000	0.0007	7	1.0000	0.0002

Scene / Group = Easy / Inexperienced			Scene / Group = Difficult / Inexperienced		
Participant	ζ	p	**Participant**	ζ	p
1	0.8500	0.0071	1	0.9500	0.0004
2	0.9500	0.0004	2	1.0000	0.0002
3	0.9500	0.0004	3	1.0000	0.0002
4	0.9000	0.0007	4	0.9500	0.0004
5	1.0000	0.0002	5	1.0000	0.0002
6	1.0000	0.0002	6	0.9500	0.0004
7	0.9500	0.0004	7	1.0000	0.0002

By computing Pearson correlation coefficients, we assessed the relationship between the preference scores of the paired comparisons and the magnitude estimation values, separately, with the performance-based measures, response time and percent error, separately. Figure 5 shows the scatter plots of the data obtained from the paired comparisons plotted against response time and percent error. Along the x-axis is the average response time and percent error for all the images used in the objective performance-based test. The paired comparison data is shown as the average number of total *votes* across participant, where in each participant, the total vote is the sum of occurrences where one algorithm condition was chosen over another. Figure 5 also shows scatter plots of the average magnitude estimation values across participants against response time and percent error. An analysis of variance concluded that the groups were not statistically significantly different from one another; thus, the values in the subjective measures are averages across the experienced and inexperienced groups. In each plot, letters that represent the algorithm conditions are superimposed on the graph point they represent and the regression lines for each data set are drawn. The letters and the algorithm conditions they represent are: O = non-degraded, C = RRF, R = MSR, A = AutoLevels, B = BBFE, H = Histogram Equalization, D = degraded, P = POSHE.

In Figure 5, the r^2 value and the p-value for the correlation coefficient for each group x scene condition are shown in the upper right-hand corner of each plot. For the difficult scene, there were significant correlations, $p \leq 0.0109$. As response time and percent error increase, we observe higher magnitude estimation values and higher number of votes in paired comparisons. Although the trend was the same for the easy scene, there were no significant correlations, $p > 0.1105$.

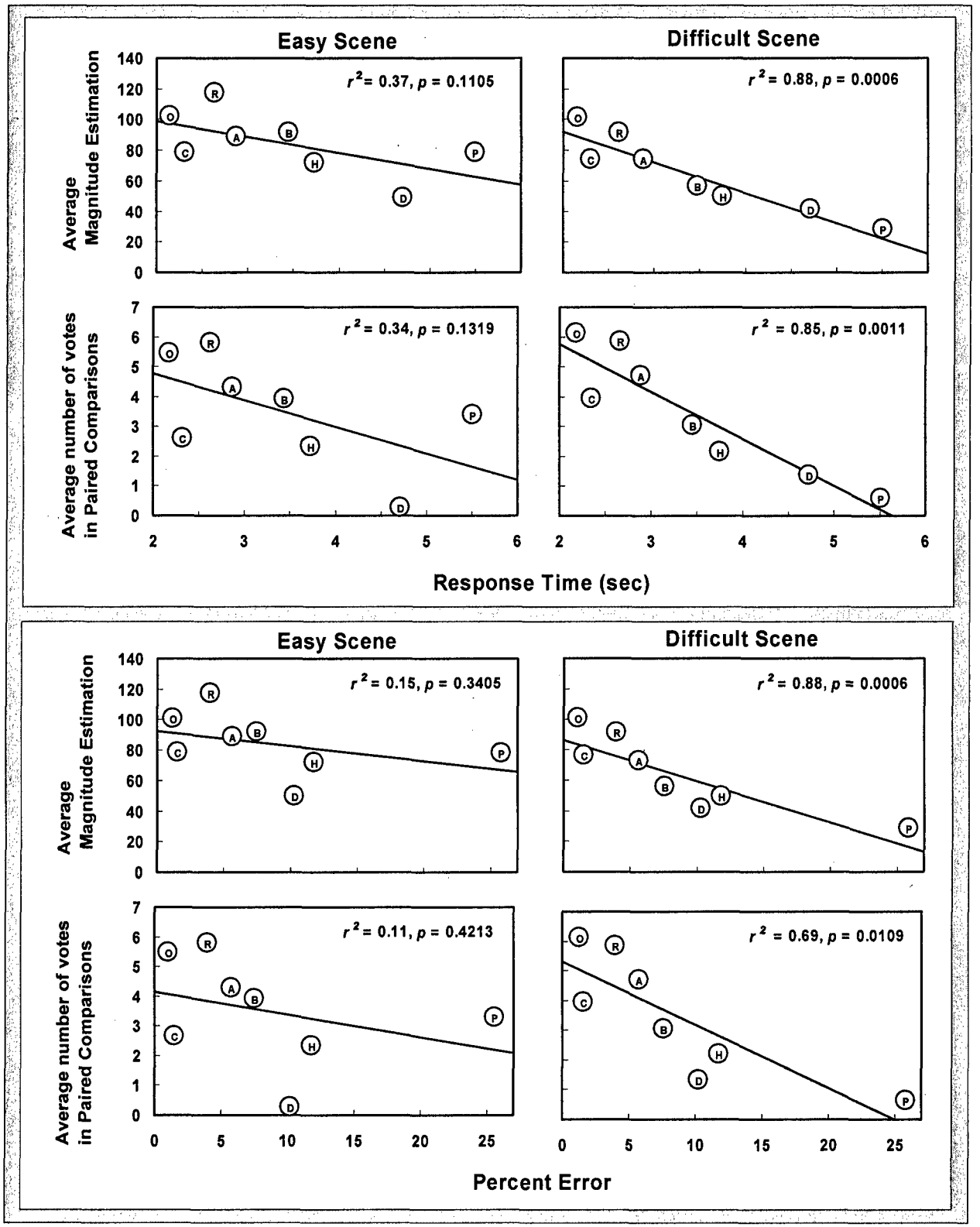

Figure 5. Scatter plots of the average number of votes in the paired comparisons and the magnitude estimation values plotted against response time and percent error.

4. DISCUSSION & CONCLUSIONS

Our objective in this study was to explore the relationship between subjective measures and objective performance-based measures. We analyzed psychometric measures as obtained through the methods of paired comparisons and magnitude estimation. Our study was inspired by the research that has been accomplished in image quality.

Our results are by no means conclusive; however, they do show interesting patterns and unfasten more questions for investigation. Firstly, we observe the difference between the difficult and easy scenes. We are not sure what contributes to one scene being more difficult than the other. It could be the spatial dynamics of the trees, roads, rivers, the location of the target, its heading, etc. Whatever the reason may be, we just know that is it more difficult to find the target in the difficult scene. To understand this difference, multiple scenes of various response times should be tested with psychometric methods. Only then may one be able to identify if certain characteristics in the scene affected the results. It may be that these psychometric methods are only appropriate for more difficult and complex imagery.

With careful inspection, it can be seen that the points "C" and "P," RRF and POSHE, respectively, may be restricting significant correlations in the easy scene (i.e., the points are pulling the regression line toward a 0 slope). Next, we observe these two points and speculate on their positions in the easy scene condition. Why would the POSHE, the worst-ranked performer in response time and percent error, have higher subjective scores than any other algorithm? Why does the RRF, the second-best performer, rank 6th in subjective measurement. One might speculate that the stark contrast levels in the POSHE give participants a sense of confidence in finding the target. For the RRF, the case may be that the "haziness" of the image does the opposite—it gives participants a sense that they may not be able to see the target when in fact they see it very well when put to the test.

We conclude that this study shows some promise for down-selecting algorithms by way of subjective assessment. At the same time, however, we must clearly state that these subjective assessments should not be used as substitutes for objective performance-based measures. This is stressed in the evidence that was presented in this study. Inconsistencies may arise as shown in POSHE and RRF. Contextual variables may play a part as shown in the differences between the easy and hard scenes. Other subjective methods may provide a better understanding of the associations between subjective and objective performance-based measures. For example, in the paired comparisons, the pairs can be presented simultaneously or the presentation can contain three or four images and the participant chooses one among the set.[7,8] Regardless of the psychometric method used, subjective measures should be used as supporting tools for objective performance-based measures which are ultimately needed for evaluation in research.

ACKNOWLEDGEMENTS

The Air Force Research Laboratory (AFRL) funded this work. The authors would like to thank each of the participants for volunteering their time to accomplish this study.

REFERENCES

1. Z. Rahman, D.J. Jobson, G.A. Woodell, and G.D. Hines, "Mutil-sensor fusion and enhancement using the Retinex image enhancement algorithm." in *Visual Information Processing XI,* Z. Rahman, R.A. Schowendgert, and S.E. Reichenbach, ed., *Proc. SPIE* 4736, 2002.
2. K.E. Neriani, T.J. Herbranson, A.R. Pinkus, C.M. Task, and H.L. Task, "Visual performance-based enhancement methodology: an investigation of three Retinex algorithms." in *Enhanced and Synthetic Vision,* J.G. Verly, ed., *Proc. SPIE 5802,* 2005.
3. Z. Wang and A.C. Bovik,"A Universal Image Quality Index,"in *IEEE Signal Processing Letters,* Vol.9,No.3, 2002.
4. J.C. Leachtenauer, "Objective quality measures assessment." in *Visual Information Processing XI,* Z. Rahman, R.A. Schowendgert, and S.E. Reichenbach, ed., *Proc. SPIE* 4736, 2002.
5. K. E. Neriani, T.J. Herbranson, G.A. Reis, A. R. Pinkus, and C. Goodyear, "Visual performance-based image enhancement methodology: an investigation of contrast enhancement. To be presented in *Enhanced and Synthetic Vision, Proc. SPIE,* 2006.
6. M.G., Kendall, and B. Babington-Smith, On the method of paired comparisons. *Biometrika, 31,* 324-345, 1939.
7. H.A., David, *The method of paired comparisons.* New York: Hafner Publishing Company., 1963.
8. B.W., Keelan and H. Urabe, "ISO 20462, A psychophysical image measurement image quality measurement standard," in *Image Quality and System Performance,* ed. Y. Miyake and D.R. Rasmussen, *Proc. SPIE-IS&T Electronic Imaging, Vol 5294.,* 181-189, 2004.